ANELL: A Web System for Portuguese Corpora Annotation

Cristina Mota[1], and Pedro Moura[1]

[1] Linguateca, Pólo do LabEL, IST, Av. Rovisco Pais I,
1049-001 Lisboa, Portugal
{cristina, pamo}@label.ist.utl.pt
http://label.ist.utl.pt and http://www.linguateca.pt

Abstract. In this paper, we briefly describe a system for annotating corpora via web which offers two operating modes: a full automatic mode and a supervised mode. The linguistic analysis of the corpora is performed by the INTEX system using the LabEL linguistic resources. The motivation, the architecture and some examples of its behavior are presented, with special emphasis on the several output formats it allows.

1 Introduction

It is not difficult to find a researcher studying or working with natural language who has not yet wanted to syntactically annotate his own corpus without having to copyright clear it, wished that his particular kind of corpus had been catered for by the (few) Portuguese corpus providers there are or thought about installing a general public tagger to handle his own text but was appalled by the technical details to overcome.

The Web system ANELL - *Anotador Electrónico do LabEL/Linguateca* contributes to the resolution of these problems. Instead of a Web interface which gives access to public corpora encoded with a common format, as it happens with the AC/DC project [1], we offer a Web system that allows the linguistic annotation of corpora that may not be public (hence may not be distributed) encoded with a common format. This approach has the advantage of giving the opportunity to different researchers who, for some reason, do not have access, the skills or the time to use (and configure) a natural language processing system, linguistically analyze and annotate their corpora using the same annotation system. INTEX [2] is the natural language processing system used to linguistically analyze the corpora and the linguistic resources were constructed by the Laboratório de Engenharia da Linguagem (LabEL).

2 System Presentation

As previously mentioned, ANELL system allows the annotation of Portuguese corpora via Web offering two main modes: one purely automatic, and the other, the *supervised* mode, requiring the intervention of a linguist who reviews the results before the annotated corpora is sent back to the user. In both modules the user may select (i) the annotation level; (ii) the linguistic resources to be applied and (iii) the format of the annotations.

The automatic mode was devised (i) to the user who wants to annotate a small corpus or needs an instant result, and (ii) so the user can try the system, selecting the most appropriate options, before sending a corpus through the supervised mode. The text to be annotated is directly typed in by the user and the resulting annotated text is presented instantly. Even though this mode is intended for small examples, the user may annotate a text of about 46 000 words in about 5 minutes. The time lost in the process is essentially due to the data transfer between the user and the web server, and other data preparation, since INTEX processes the text in a few seconds.

The supervised mode allows the reviewing of the results by a linguist, before the annotated text is sent back to the user, and makes possible to annotate large corpora (with millions of words). Both the corpora and the resulting annotated corpora files are sent either in plain text format or any another format chosen by the user (html, doc, pdf, etc.). Considering that the corpus may be of a considerable size, when the result is ready, an automatic e-mail message is sent to the user indicating the URL location where the annotated corpora can be downloaded.

The diagram in Figure 1 illustrates the general architecture of the system. The automatic server receives and processes the requests related with the automatic mode, functioning as a web simplified interface of INTEX. Additionally, it converts the text analyzed by INTEX which is represented by finite-state transducers into the annotation format chosen by the user. After receiving the data to be processed there is no human interaction until the annotated corpus is viewed by the user. The supervised server has two main modules: one communicates with the user and the other with the linguist. The module that communicates with the user receives the request which will be put on an offline waiting list. This offline list is accessed by the linguist who will review the text analyzed by INTEX before the final annotated text is created and sent to the user. For the moment, the linguist can only verify the parameters provided by the user and instruct the supervised server to resume the annotation request. Consequently, the supervised server will configure INTEX to process the text. After receiving the result it will convert the analyzed text into the annotation format specified by the user and notify the user by e-mail of the process conclusion.

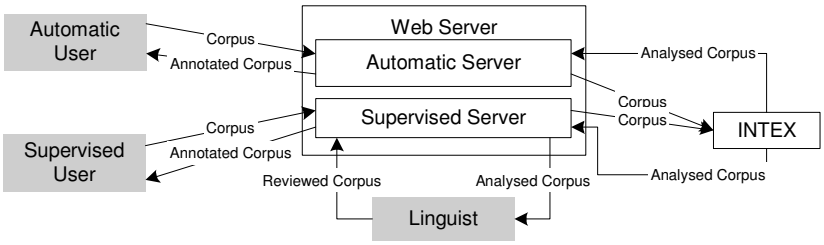

Figure 1 - ANELL general architecture

At this time, the service is installed in a Windows XP 5.1 running Apache 1.3 and INTEX 4.23e. The CGI are implemented in Clisp 2.30.

3 The Linguistic Analysis

As previously mentioned, the corpora are analyzed by INTEX using the LabEL linguistic resources.

INTEX is a modular natural language processing system based on finite-state technology which allows high levels of data compression and fast processing, without loss of accuracy in the results. This system is comprised of several command line tools that are also used by a windows interface which communicates with those tools to perform various NLP operations in an integrated environment. This system allows, on the one hand, to construct large-scale linguist resources (dictionaries and grammars), and, on the other hand, to apply those resources to large corpora in order to perform different text analysis tasks: text indexing, location of morpho-syntactic patterns and multi-word expressions, lexical disambiguation and tagging, parsing, among others.

The linguistic data used by the INTEX system were constructed by the LabEL team [3], [4] and are all internally represented by finite-state transducers. At present, the ANELL user may opt for applying: (i) simple word dictionaries (*árvore* (tree), *desagradável* (unpleasant)); (ii) compound word dictionaries (*carta aberta*, *porta-chaves* (nouns), *de qualquer maneira* (adverb)); (iii) dictionaries of acronyms (*ONU*) and their lexical base forms (*Organização das Nações Unidas*); (iv) disambiguation grammars of noun phrases containing adjectives.

4 The Annotation

The result of the analysis produced by INTEX is a sequence of transducers, each representing the various lexical analyses of the words within one sentence of the text. It is based on these transducers that the system will create the final annotated text.

Currently, there are four different types of annotation produced by the system which may have different levels of annotation (lemma, grammatical category, syntactic subclassification, semantic attributes and morphological attributes). These formats are still being adjusted and they can be changed easily.

We will illustrate the available annotation formats[1] based on the analysis of the sentence: *Isso é um barril de pólvora* (That is a barrel of gunpowder).

Text with indexed parentheses. Each new sentence identified by INTEX starts (and ends) with a parenthesis indexed with number *1* representing the concatenation of the words of the sentence. Whenever a lexical unit is ambiguous, the various lexical ambiguities are enclosed between parentheses indexed with the previous number incremented by one.

(1 *{isso(isso):PRO:Dem} {é(ser:Y2s,P3s,P4s,P2s):V} {um(um:ms):DET:Art, Ind}* ***(2 (3*** *{barril(barril:ms):N} {de(de):PREP} {pólvora(pólvora:fs):N}* ***3)*** *{barril de pólvora(barril de pólvora:ms):N}* ***2) 1)***

Text with associated type. Instead of using indexed parentheses to group the lexical units and the lexical ambiguities of those units, as it happens in the previous case, the parentheses begin with the operator OR or AND, depending on the grouping type.

(:AND *{isso(isso):PRO:Dem} {é(ser:Y2s,P3s,P4s,P2s):V} {um(um:ms):DET:Art, Ind}* ***(:OR (:AND*** *{barril(barril:ms):N} {de(de):PREP} {pólvora(pólvora:fs):N})* *{barril de pólvora(barril de pólvora:ms):N}****))***

Indented text with associated type. This format is similar to the previous one, but each different analysis is in a new line, indented according to the depth of the analysis.

Indented one word per line. This is the simplest format. It is identical to the preceding format but there are neither parentheses nor operators.

5 Final remarks

The presented system is already working (http://label.ist.utl.pt/pt/anell.html or http://www.linguateca.pt/anell.html) and we plan to make a public announcement soon; we believe it can already be extremely useful in its present form. Nevertheless, we are developing two major envisaged functionalities: one is the integration of reviewing tools to aid the linguist on the reviewing process before sending the final annotated text to the user; the other is to give the possibility of annotating corpora that have been previously marked up. As soon as the system starts being used, we intend to assess its adequacy in real use contexts.

Acknowledgements

We are grateful to Diana Santos, Elisabete Ranchhod and Paula Carvalho for their support and valuable suggestions throughout the preparation of this paper.

[1] In order to simplify the illustration of the annotation formats, we decided to use only the lemma and syntactic category annotation levels. All the linguistic resources were applied but the disambiguation grammars (so the ambiguities annotation can be seen in).

References

1. Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavriladou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.): Proceedings of LREC2000, Athens (2000) 205-210
2. Silberztein, M.: Dictionnaires électroniques et analyse automatique de textes: le système INTEX. Masson Ed., Paris (1993)
3. Ranchhod, E.: Ressources linguistiques du portugais implémentées sous INTEX. In: Fairon, C. (ed.): Analyse Lexicale et Syntaxique: Le système INTEX. Lingvisticae Investigationes, XXII. John Benjamins Publishing Company, Amsterdam/Philadelphia (1998-1999) 263-278
4. Eleutério, S., Ranchhod, E., Mota, C., Carvalho, P.: Dicionários Electrónicos do Português. Características e Aplicações. In: Miyares, L., Moreno C., Silva, M. (eds.): Actas do VIII Simposio Internacional de Comunicación Social, Vol. 1, Centro de Lingüística Aplicada, Santiago de Cuba (2003) 636-642